THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Traffic Measurement Biases Induced by Partial Sampling

By A. DESCLOUX

(Manuscript received March 22, 1973)

Under equilibrium conditions, the sample average of the delays encountered by all the calls submitted during a given time interval is an unbiased estimate of the mean of the delay distribution. If some of the delays are not observed, the resulting sample average need no longer be an unbiased estimator of the corresponding population mean. This is the case when, for instance, only a limited number of delays can be timed simultaneously. The purpose of this paper is to investigate these biases for queuing systems when only one clock is available and thus one delay only can be measured at a time. It is shown that, regardless of the order of service, the expected value of the observed average delays is always smaller than the mean waiting time for all calls.

Although the average delay on all calls is independent of the order of service, the measurement biases resulting when only one delay can be measured at once depend on the queue discipline. In particular, we shall show that the average delay for all calls is always larger than the average delay of the observed calls even if these calls are always served last (observed-call served-last).

I. INTRODUCTION

The following remarks due to J. F. C. Kingman appear in the *Proceedings of the Symposium on Congestion Theory* held at the University of North Carolina in 1964 (Ref. 1, pp. 314–315): "To illustrate the pitfalls of inference from congestion systems, let me tell a (more or less true) story. It was desired to estimate the mean waiting time in a particular queuing system, and for technical reasons, only one customer could be timed at once. Thus the waiting time ω_1 of a customer was measured. When he entered service, the next customer to arrive was observed and his waiting time ω_2 was noted. This procedure continued, the waiting times ω_3, ω_4, $\cdots$ being measured, and, for

large n,

$$n^{-1}(\omega_1 + \omega_2 + \cdots + \omega_n)$$

was used as an estimate of the waiting time. It is, however, strongly biased and inconsistent, because of the selection of the customers to he observed. The mean waiting time is overestimated by a factor which becomes arbitrarily large as the traffic intensity approaches one." We stress that, according to this sampling procedure, a customer is ohserved if and only if it arrives when the clock is free.

Another instance of biases induced by the measurement procedure is reported by Oberer and Riesz.[2] These authors have investigated the possihility of estimating hlocking prohahilities in telephone networks by means of test calls generated by a single source repeatedly calling a dedicated numher. Their study shows that the proportion of hlocked test-calls does not yield a suitable estimate of the grade of service as it is markedly biased downwards. As expected the bias becomes larger as the intervals hetween consecutive (nonoverlapping) test-calls becomes smaller. It is also established in Ref. 2 that the relative test-call biases increase as the blocking probability decreases.

It is worth noting that the hiases studied in Ref. 2, as well as here, are of a different sign than those referred to by Kingman. Much more important, however, is the fact that measurement techniques which, superficially, appear to be adequate may prove to be very unreliable. It is thus becoming increasingly clear that great care is required in the design of performance measurements for stochastic service systems so that unanticipated biases are not encountered.

The purpose of this paper is to investigate the effects of partial sampling on the estimate of the mean (overall) waiting time obtained by averaging measured delays. The biases induced by such limitations will be studied here for M/G/1 and GI/M/s when at most one call can be observed at once and the estimation procedure is as described by Kingman. We shall see that, in these systems, the equilihrium average delay of the observed calls is always smaller than the equilibrium average delay for all calls.

It is well known that the average delay for all calls is the same for all queue disciplines which are independent of the lengths of the individual calls (no other type of queue disciplines will he considered here). As we shall see, this is not true of the mean *measured* delay when only one delay can be recorded at a time. In this case, both the unconditional and the conditional average delays† are (as expected)

† As customary, unconditional and conditional delays pertain to arbitrary and delayed calls respectively.

smallest when the observed calls are served first and largest when they are served last. (In particular, the second extreme case occurs in systems with first-come last-served queue discipline.) Furthermore, in view of the general inequality mentioned at the end of the preceding paragraph, the upper bound for the unconditional average delay of the observed calls (which is reached when the observed calls are served last) is always a strict lower bound for the unconditional average delay for all calls!

The preceding result pertains to unconditional delays and does not always hold for the average delay of those sampled calls which encounter a delay. Thus for M/M/s, the conditional average delay for all delayed calls is equal to the conditional average delay of the observed delayed calls so long as these are always served last. (Note that for M/M/s, the average delay of the delayed calls is equal to the average length of the busy period, and that the waiting-time distribution of the observed calls coincides with the busy-period distribution for the observed-served-last measurement procedure.) In contrast, for the M/Γ_k/1 queue, the upper bound for the conditional average delay of the observed delay calls is larger than the average conditional delay for all delayed calls when $k > 1$, the inequality being reversed whenever $0 < k < 1$. (Γ_k is used here to designate the gamma distribution with mean 1 and variance k^{-1}. Thus M/Γ_1/s is identical to M/M/s. When k is an integer, Γ_k is the Erlangian distribution often designated E_k.)

Expressions for the moments of the equilibrium delay-distribution of the observed calls are given for M/G/1 and first-come first-served queue discipline. The equilibrium delay-distribution of the observed calls is also derived for M/M/s with order-of-arrival service. (Corresponding results for the "observed-call served-first" and the "observed-call served-last" measurement procedures are immediate.) These formulas are used to show that the biases induced by partial sampling can be quite substantial.

When the average service-time is unity, an assumption made throughout, the average delay, EW, for the single-server queue M/G/1 is given by the formula (Ref. 3, pp. 46–50):

$$EW = \alpha m_2/2(1 - \alpha),$$

where α is the server occupancy and m_2 is the second moment about 0 of the service-time distribution. Since m_2 can be arbitrarily large, no bound can be placed on the value of EW. But when only one delay can be timed at once, we shall see that the expectation of the observed delays cannot exceed $1/(1 - \alpha)$. Therefore for any prescribed value

of α, it is always possible to find service-time distributions for which the ratio of the average delay for all calls to the average delay of the observed calls exceeds any given bound.

To simplify the exposition we restrict ourselves to full-access delay systems with recurrent inputs in which delays are measured by means of a single clock. Some of the results obtained below can, however, be extended to more general situations.

(In the sequel, W, with or without affix, is used to designate the waiting time of an arbitrary call while W_*, with or without affix, is used as the generic symbol for the observed delays when only one clock is available.)

II. A GENERAL DELAY FORMULA

Consider the queuing system GI/G/s and suppose that the arrival and service-time distributions are such that equilibrium can be reached. (To avoid trivial qualifications we assume throughout that the mean interarrival time is finite. For the same reason, the underlying distributions are also supposed to be such that simultaneous occurrences of events need not be considered.) The purpose of this section is to derive a formula relating the average delay of the observed calls to the equilibrium probability, Φ, that an observed call has immediate access to a server. To this end we prove first that

$$\Phi = (1 - B)(1 + A)/[(1 - B)(1 + A) + B], \tag{1}$$

where B is the equilibrium blocking probability for all calls and A is the expectation of the number of unobserved calls originating during the waiting time of an arbitrary observed delayed call.

It follows from (1) that the probability that an observed call is blocked is always (strictly) smaller than the overall probability of delay (so long as $B \neq 0$ or $B \neq 1$, two trivial cases that we exclude from our considerations). This, of course, is a consequence of the fact that all nonblocked calls are observed whereas, with one clock only, some delays may not be recorded.

We turn now to the proof of (1). Consider an infinite sequence of consecutive calls and for the ith call ($i = 1, 2, \cdots$) let

$$X_i = \begin{cases} 0 \text{ if the } i\text{th call is delayed,} \\ 1 \text{ if the } i\text{th call is not delayed,} \end{cases}$$

$$Y_i = \begin{cases} 0 \text{ if the } i\text{th call is not observed,} \\ 0 \text{ if the } i\text{th call is observed and not delayed,} \\ 1 \text{ if the } i\text{th call is observed and delayed.} \end{cases}$$

Let $\epsilon > 0$. Then assuming that the system is in equilibrium when the first call arrives, we have, by the integral stationarity theorem (Ref. 4, p. 419),

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{(X_1 + Y_1 + \epsilon) + \cdots + (X_n + Y_n + \epsilon)} = \frac{EX_1}{E(X_1 + Y_1) + \epsilon} \quad (2)$$

and

$$\lim_{n\to\infty} \frac{(X_1 + Y_1) + \cdots + (X_n + Y_n)}{(X_1 + \epsilon) + \cdots + (X_n + \epsilon)} = \frac{E(X_1 + Y_1)}{EX_1 + \epsilon}, \quad (3)$$

with probability 1. However, since (Ref. 4, p. 421)

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = EX_1 = \Pr[X_1 = 1] > 0,$$

with probability 1, there is, for almost all realizations of the process, an integer n such that the ratios

$$\frac{X_1 + \cdots + X_m}{(X_1 + Y_1) + \cdots + (X_m + Y_m)}, \qquad m \geqq n,$$

are well defined. Hence, by (2) and (3), we have

$$\frac{EX_1}{E(X_1 + Y_1) + \epsilon} \leqq \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{(X_1 + Y_1) + \cdots + (X_n + Y_n)} \leqq \frac{EX_1 + \epsilon}{E(X_1 + Y_1)},$$

with probability 1 and, letting ϵ tend to 0,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{(X_1 + Y_1) + \cdots + (X_n + Y_n)} = \frac{EX_1}{E(X_1 + Y_1)}, \quad (4)$$

with probability 1.

[The preceding derivation makes use of the fact that the stationarity—and hence the integral stationarity—of the processes $\{X_i, i = 1, \cdots\}$ and $\{X_i + Y_i, i = 1, \cdots\}$ follows from the property that the random variables X_i and Y_i, $i = 1, \cdots$, whose means are finite, are "translates" defined on the stationarity queuing process (Ref. 4, p. 417 ff.).

[The formulas in Ref. 4, p. 419, Theorem A, involve conditional expectations with respect to fields of invariant events. Under the present circumstances these expressions can he simplified. Indeed let us specify the state of the system, $\mathcal{T}_t$ at time t, by means of the vector whose components are the arrival time of the last request placed before t and the elapsed portions of the service-times in progress

at time t. Then the only invariant sets (Refs. 4 and 5) of the process $\{\mathcal{T}_t, -\infty < t < \infty\}$ are the whole space and the null-set. This property, in turn, implies that the conditional expectations of the random variables X_i and $X_i + Y_i$ relative to their invariant fields can be replaced by the unconditional expectations EX_i and $E(X_i + Y_i)$, respectively. These and other similar substitutions are made here without formal justification.]

Consider now an infinite sequence of observed calls and let

$$Z_i = \begin{cases} 0 \text{ if the } i\text{th observed call is delayed,} \\ 1 \text{ if the } i\text{th observed call is not delayed.} \end{cases}$$

Then (Ref. 4, p. 421)

$$\lim_{n\to\infty} \frac{Z_1 + Z_2 + \cdots + Z_n}{n} = EZ_1 = \Phi. \tag{5}$$

Since (4) and (5) are both equal to the proportion of observed calls with zero delay over an interval of infinite length, we have

$$\Phi = EX_1/E(X_1 + Y_1). \tag{6}$$

We note that EX_1 is equal to the probability that a call (observed or not) is not delayed. Hence

$$EX_1 = 1 - B, \tag{7}$$

and to complete the proof of (1) we have to show that

$$EY_1 = B/(1 + A). \tag{8}$$

To this end consider again a stationary sequence of observed calls and let A_i be the number of unobserved calls placed during the waiting time of the ith call that is both observed and delayed. Then we have (Ref. 4, pp. 419–421)

$$\lim_{n\to\infty} \frac{n}{n + A_1 + \cdots + A_n} = \frac{1}{1 + EA_1} = \frac{1}{1 + A}, \tag{9}$$

with probability 1.

Furthermore, by the integral stationarity theorem (Ref. 4, p. 419) and a simple ϵ-argument of the type used in the proof of (2), we have:

$$\lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{(1 - X_1) + \cdots + (1 - X_n)} = \frac{EY_1}{E(1 - X_1)}, \tag{10}$$

with probability 1.

Since the left-hand sides of (9) and (10) are both equal to the proportion of delayed calls that are observed, we have

$$EY_1 = E(1 - X_1)/(1 + A) = B/(1 + A).$$

This completes the proof of (1).

Our next step will now be to relate A to the average delay, EW_*, of the observed calls. Let W_{i*} be the delay of the ith observed call and let U_i be the interval between the end of the ith and the beginning of the $(i + 1)$st measurement. Let also I_n be the interval between the arrival epochs of the nth and $(n + 1)$st call (in the whole sequence of calls, observed or not). Then we have:

$$(W_{1*} + U_1) + \cdots + (W_{n*} + U_n) = I_1 + \cdots + I_{K_n}, \tag{11}$$

where K_n, a random variable, is equal to the number of calls placed during the interval that starts with an observed call and ends just before the beginning of the $(n + 1)$st measurement. By the stationarity theorem, we have (Ref. 4, p. 421):

$$\lim_{n\to\infty} \frac{(W_{1*} + U_1) + \cdots + (W_{n*} + U_n)}{n} = E(W_{1*} + U_1), \tag{12}$$

with probability 1, and, by the strong law of large numbers (note that the I_n's are, by assumption, independent random variables with finite means and that $K_n \geqq n$),

$$\lim_{n\to\infty} \frac{I_1 + I_2 + \cdots + I_{K_n}}{K_n} = \alpha^{-1}, \tag{13}$$

with probability 1, where α^{-1} is the expected interarrival-time.

Furthermore,

$$K_n = [Z_1 + (1 - Z_1)(1 + A_1)] + \cdots + [Z_n + (1 - Z_n)(1 + A_n)],$$

so that

$$\lim_{n\to\infty} \frac{K_n}{n} = E[Z_1 + (1 - Z_1)(1 + A_1)] = \Phi + (1 - \Phi)(1 + A), \tag{14}$$

with probability 1.

Combining (11)–(14) we find that

$$\alpha E(W_{1*} + U_1) = 1 + A(1 - \Phi). \tag{15}$$

In particular, when the input is Poissonian, $EU_1 = \alpha^{-1}$ and (15) reduces to

$$\alpha E W_{1*} = A(1 - \Phi). \tag{16}$$

Thus, taking (1) into account, we find that:

$$EW_* \equiv EW_{1*} = (\Phi + B - 1)/\alpha(1 - B). \tag{17}$$

It should be noted that the preceding relation is valid regardless of the order of service.

When the calls are served in order of arrival, (16) is an immediate consequence of the fact that the waiting time of any given call is not affected hy thc stream of requests placed after its arrival epoch. This is also true when the observed calls are always served first and (16) can then be written down with equal ease.

We note that (17) can he ohtained quickly whenever the epochs at which measurements begin or terminate constitute a renewal process. In such cases, the expected numher of ohservations in time t is (asymptotically)

$$(EW_* + \alpha^{-1})^{-1} \cdot t + 0(1),\ t \text{ large}, \tag{18}$$

and the expected number of arrival points at which the system is empty in time t is

$$\alpha(1 - B) \cdot t + 0(1),\ t \text{ large}. \tag{19}$$

The long-term proportion of ohserved calls with no delay is given hy the ratio of (19) to (18) with probahility 1 (cf. Ref. 6, p. 264, alternative form of Theorem IV):

$$\alpha(1 - B)(EW_* + \alpha^{-1}). \tag{20}$$

Since the prohability Φ that an arbitrary call is not delayed is independent of the past, a simple application of the strong law of large numhers show that (20) may he equated to Φ and (17) therefore holds. An instance where the preceding argument can he applied is the M/G/1 system with ohserved calls always served last. With this order of service, there is exactly one call in the system at the termination of each measurement. These epochs constitute a renewal process since they also coincide with the beginnings of the service-times of the observed calls. With an obvious change, the previous argument remains true for M/M/s with observed-calls served-last.

III. TWO EXTREME CASES

In this section we show that if only one clock is availahle then the expected average delay of the ohserved calls is largest when the observed calls are served last and smallest when the ohserved calls

are served first. These relations are not statistical: They are satisfied by all the realizations of the process over any finite or infinite time interval regardless of the arrival and service-time distributions. (To avoid ambiguities, we assume that the timing device is free at the heginning of the realizations.)

We note first that under any measurement procedure, all the calls which arrive when all the servers are husy, hut no request is waiting, are observed. These calls are the only delayed calls that are observed when the observed calls are served last. Therefore (*i*) the number of observed delayed calls takes its smallest value for the observed-served-last procedure and (*ii*) during the measurement of a delay, D, under this particular procedure any ohserved delay under any alternate single-clock measurement procedure cannot exceed D. Comhining these two facts and taking into account that all calls with zero delay are ohserved we may conclude that the ohserved average delay takes always its largest value when the ohserved calls are served last, as is the case for the first-come last-served queue discipline.

When the observed calls are served first, we note that (*i*) the number of observed delays over any busy period (initial or not) is never smaller than for any other single-clock measurement procedure and (*ii*) to each ohserved delayed call there corresponds, under any other single-clock measurement procedure, one call whose delay is at least as large and this correspondence involves all the ohserved delays under the alternate procedure. All calls with zero delay are again ohserved and the average delay of the ohserved calls takes therefore its smallest value when the ohserved calls are served first.

Clearly the conditional average delays of the observed calls do have the same property.

IV. Bounds for the average delay of the observed calls in M/G/1

The object of this section is to determine the upper and lower bounds for the average delay, EW_*, of the ohserved calls in M/G/1. These hounds, as we have seen, are reached when the ohserved calls are served last and first respectively. Under the present conditions, formula (17) may be written as follows:

$$EW_* = (\Phi + \alpha - 1)/\alpha(1 - \alpha). \tag{21}$$

For a given server occupancy α, EW_* is a monotone increasing function of $\Phi = \Phi(\alpha)$. Since $\Phi < 1$, (21) implies that $EW_* < (1 - \alpha)$. Hence for $\alpha < 1$, EW_* is always bounded (hut EW is not).

It will be convenient to define here the service backlog, at a given instant t, as the sum of the service-times of all waiting requests plus the residual of the service-time of the request being served. [When calls are served in order of arrival, the service backlog is equal to the virtual waiting time (Ref. 3, p. 59 ff.).]

Now let $F(\cdot)$ be the stationary cumulative distribution of the service backlog at the end of a measurement. The probability, $\Phi(\alpha)$, that an observed call does not suffer a delay is simply the Laplace-Stieltjes transform of $F(\cdot)$ evaluated at α since it is equal to the probability that no call originates during a time interval whose length is that of the service backlog:

$$\Phi(\alpha) = \int_0^\infty e^{-\alpha t} dF(t).$$

Writing $\sigma(\cdot)$ for the Laplace-Stieltjes transform of the service-time distribution we have the following inequality:

$$\Phi(\alpha) \leqq \sigma(\alpha). \tag{22}$$

This inequality is a consequence of the fact that, at the conclusion of a measurement, the service backlog may be represented as the sum of two independent random variables, one of which is the full service-time of the request whose delay has just come to an end while the other is equal to the sum of the service-times of all the waiting requests. Writing $R(\cdot)$ for the c.d.f. of the latter and $S(\cdot)$ for the service-time distribution, we have:

$$\begin{aligned}\Phi(\alpha) &= \int_0^\infty e^{-\alpha t} d \int_0^t R(t - v) dS(v) \\ &= \sigma(\alpha) \int_0^\infty e^{-\alpha t} dR(t) \leqq \sigma(\alpha).\end{aligned}$$

When the observed calls are served last,

$$R(t) = \begin{cases} 1 & \text{for} \quad t \geqq 0, \\ 0 & \text{for} \quad t < 0, \end{cases}$$

and

$$\Phi(\alpha) = \sigma(\alpha).$$

We can therefore conclude that

$$EW_* \leqq \frac{\sigma(\alpha) + \alpha - 1}{\alpha(1 - \alpha)}. \tag{23}$$

We are now in a position to prove that the average delay, EW_*, is always smaller than the average delay for all calls (observed or not). Since the service-time is unity, we have:

$$\Phi(\alpha) \leqq \sigma(\alpha) = \int_0^\infty e^{-\alpha t} dS(t) < 1 - \alpha + \frac{\alpha^2 m_2}{2},$$

where m_2 is the second moment of S about the origin. Hence, substituting $1 - \alpha + \alpha^2 m_2/2$ for $\sigma(\alpha)$ in (23) we find that, irrespective of the service order:

$$EW_* < \frac{\alpha m_2}{2(1-\alpha)} = EW. \tag{24}$$

By (23) and the equality in (24) we also have:

$$EW/EW_* > \frac{\alpha m_2}{2},$$

so that, for any given α, we can always find a service-time distribution such that EW/EW_* exceeds any preassigned value.

We now derive an absolute lower bound for EW_*. As shown above, this bound can be found by assuming that the observed calls are served first. Our first step here is to determine A.

For the observed-served-first procedure, the circumstances under which a positive delay can be observed are as follows: at some time the clock is not in use and a service-time begins, and during this service-time a new call arrives. Thus A is the conditional expectation of the number of arrivals minus 1 during an arbitrary service-time given that at least one call is placed during a service-time. A, therefore, is given by the formula

$$\begin{aligned} A &= \int_0^\infty \sum_{n=1}^\infty (n-1)\frac{(\alpha t)^n}{n!} e^{-\alpha t} dS(t) \Big/ \int_0^\infty (1 - e^{-\alpha t}) dS(t) \\ &= \int_0^\infty [\alpha t - 1 + e^{-\alpha t}] dS(t) \Big/ \int_0^\infty (1 - e^{-\alpha t}) dS(t) \\ &= [\alpha - 1 + \sigma(\alpha)]/[1 - \sigma(\alpha)]. \end{aligned}$$

By means of (1) and (21), it is now readily shown that, for the observed-calls-served-first procedure:

$$\Phi(\alpha) = \frac{(1-\alpha)}{2 - \alpha - \sigma(\alpha)}$$

and

$$EW_* = \frac{\sigma(\alpha) + \alpha - 1}{\alpha[2 - \alpha - \sigma(\alpha)]}.$$

Summing up, we have the following inequalities for Φ and EW_*, regardless of the measurement procedure:

$$\frac{1 - \alpha}{2 - \alpha - \sigma(\alpha)} \leqq \Phi(\alpha) \leqq \sigma(\alpha), \tag{25}$$

and

$$\frac{\sigma(\alpha) + \alpha - 1}{\alpha[2 - \alpha - \sigma(\alpha)]} \leqq EW_* \leqq \frac{\sigma(\alpha) + \alpha - 1}{\alpha(1 - \alpha)}. \tag{26}$$

For exponential service-times, (25) and (26) reduce to

$$(1 - \alpha^2)/(1 + \alpha - \alpha^2) \leqq \Phi_1(\alpha) \leqq 1/(1 + \alpha),$$
$$\alpha/(1 + \alpha - \alpha^2) \leqq EW_{*1} \leqq \alpha/(1 - \alpha^2).$$

(The subscript 1 is added to Φ and EW_* to indicate that the service-times are exponentially distributed.)

V. THE SINOLE-SERVER QUEUE M/G/1 WITH ORDER-OF-ARRIVAL SERVICE

In this section we consider the M/G/1 queue under the assumption that the calls are served in order of arrival. Our principal aim here is to determine $\Phi = \Phi(\alpha)$ and then, by means of (17), EW_*. To this end let p_n be the probability tbat there are n calls in the system immediately after the conclusion of a measurement. Then, relating the state probabilities at two consecutive conclusions of delay measurements, we find that

$$\begin{aligned} p_1 &= \sum_{n=1}^{\infty} p_n \int_0^{\infty} e^{-\alpha t} dS^{(n)}(t) + \sum_{n=1}^{\infty} p_n \int_0^{\infty} \alpha t e^{-\alpha t} dS^{(n)}(t), \\ p_k &= \sum_{n=1}^{\infty} p_n \int_0^{\infty} \frac{(\alpha t)^k}{k!} e^{-\alpha t} dS^{(n)}(t), \qquad k > 1, \end{aligned} \tag{27}$$

where $S^{(n)}$ is the nth convolution of the service-time distribution, S, with itself.

Now let

$$G(x) \equiv \sum_1^{\infty} p_n x^n.$$

Equations (27) yield:

$$G(x) = \sum_{m=1}^{\infty} p_m x^m = \sum_{m=1}^{\infty} x^m \sum_{n=1}^{\infty} p_n \int_0^{\infty} \frac{(\alpha t)^m}{m!} e^{-\alpha t} dS^{(n)}(t)$$

$$+ x \sum_{n=1}^{\infty} p_n \int_0^{\infty} e^{-\alpha t} dS^{(n)}(t)$$

$$= \sum_{n=1}^{\infty} p_n \int_0^{\infty} \left[\sum_{m=1}^{\infty} \frac{(\alpha t x)^m}{m!} e^{-\alpha t} \right] dS^{(n)}(t) + x \sum_{n=1}^{\infty} p_n \sigma^n(\alpha)$$

$$= \sum_{n=1}^{\infty} p_n \int_0^{\infty} e^{-\alpha t}(e^{\alpha t x} - 1) dS^{(n)}(t) + x \sum_{n=1}^{\infty} p_n \sigma^n(\alpha)$$

$$= \sum_{n=1}^{\infty} p_n \sigma^n[\alpha(1 - x)] - (1 - x) \sum_{n=1}^{\infty} p_n \sigma^n(\alpha)$$

$$= G\{\sigma[\alpha(1 - x)]\} - (1 - x) G[\sigma(\alpha)].$$

Summing up, we have the relation

$$G(x) = G\{\sigma[\alpha(1 - x)]\} - (1 - x) G[\sigma(\alpha)]. \tag{28}$$

Note also that

$$\Phi(\alpha) = \sum_{1}^{\infty} p_n \int_0^{\infty} e^{-\alpha t} dS^{(n)}(t) = G[\sigma(\alpha)].$$

Let $x_0 = \sigma(\alpha)$ and $x_n = \sigma[\alpha(1 - x_{n-1})]$, $n = 1, 2, \cdots$.

Since $0 \leqq \alpha < 1$, we have $x_0 < x_1 < \cdots \leqq 1$ and $\lim_{n \to \infty} x_n$ does therefore exist. With this notation, we obtain, from (28):

$$\begin{aligned} \Phi(\alpha) &= G(x_1) - (1 - x_0)\Phi(\alpha), \\ G(x_1) &= G(x_2) - (1 - x_1)\Phi(\alpha), \\ &\vdots \\ G(x_{n-1}) &= G(x_n) - (1 - x_{n-1})\Phi(\alpha). \end{aligned} \tag{29}$$

Adding up these relations, we find that:

$$\Phi(\alpha)\left[1 + \sum_{m=0}^{n-1} (1 - x_m)\right] = G(x_n),$$

and, by passing to the limit,

$$\Phi(\alpha)\left[1 + \sum_{m=0}^{\infty} (1 - x_m)\right] = 1. \tag{30}$$

(Note that $\lim_{n \to \infty} G(x_n)$ exists since the x_n constitute a positive monotone-increasing sequence bounded by 1. By letting $n \to \infty$ in the last of the relations (29) it follows immediately that $\lim_{n \to \infty} x_n = 1$ so

long as $\Phi(\alpha) \neq 0$. This last condition is however clearly satisfied whenever $\alpha < 1$.)

In particular, when the service-times are negative exponential, we have: $S(t) = 1 - e^{-t}$, $t \geqq 0$, $\sigma(s) = (1 + s)^{-1}$, and $(1 - x_m) = \alpha^{m+1}/(1 + \alpha + \cdots + \alpha^{m+1})$. Hence, by (30),

$$\Phi(\alpha) = \left[1 + \sum_{m=0}^{\infty} \frac{\alpha^{m+1}}{1 + \alpha + \cdots + \alpha^{m+1}}\right]^{-1}. \tag{31}$$

We examine briefly the case where the service-times have a gamma distribution with parameter k (the subscript k is added to the symbols considered earlier in order to stress their dependence on k). We have, in this case:

$$\sigma_k(\alpha) = [k/(k + \alpha)]^k.$$

Then $\sigma_k(\alpha)$ is a strictly decreasing function of $k(> 0)$ as can be seen by taking the derivative of $\ln \sigma_k^{-1}(\alpha) = \ln (1 + \alpha/k)^k$ and using the inequality $\ln (1 - x) > x/(1 + x)$, $x > 0$ (Ref. 7, p. 68). This monotonicity property of σ_k implies that

$$x_{k0} > x_{k+h,0}, \qquad x_{k1} > x_{k+h,1}, \cdots, \qquad k > 0, \qquad h > 0,$$

and we have, therefore:

$$\sum_{m=0}^{\infty} (1 - x_{km}) < \sum_{m=0}^{\infty} (1 - x_{k+h,m}), \qquad h > 0,$$

so that

$$\Phi_k(\alpha) > \Phi_{k+h}(\alpha),$$

and from (21)

$$EW_{*k} > EW_{*,k+h}, \qquad k > 0.$$

For $k = 1$ (exponential service-time) the conditional average delay for the delayed calls under the observed-last-served procedure is equal to the average length of the busy period. To see this one need only note that each positive observed delay begins with an arrival that occurs when there is exactly one customer in the system and ends when, for the first time thereafter, there is no waiting customer. Hence, for $k = 1$, the conditional delay distribution of the observed calls is the same as the busy-period distribution (Ref. 8, p. 32). Since the average length of the busy period and the average of the conditional waiting times of all the delayed calls are both equal to $(1 - \alpha)^{-1}$ (Ref. 3, p. 63), we have

$$\frac{EW_1}{\alpha} = \frac{\overline{EW}_{*1}}{1 - \sigma_1(\alpha)} = \frac{\sigma_1(\alpha) + \alpha - 1}{[1 - \sigma_1(\alpha)]\alpha(1 - \alpha)} = \frac{1}{1 - \alpha}, \tag{32}$$

where $\overline{EW}_{*1}$ designates the average delay when the observed calls are served last.

Clearly, the inequalities (22) and (23) imply that

$$\frac{EW_{*k}}{1 - \Phi_k} \leqq \frac{\sigma_k(\alpha) + \alpha - 1}{[1 - \sigma_k(\alpha)]\alpha(1 - \alpha)}. \tag{33}$$

We note that if the service-times have a gamma distribution with transform $\sigma_k(s) = (k/k + s)^k$, then the conditional average delay on all delayed calls is given by the formula (Ref. 3, p. 50):

$$\frac{EW_k}{\alpha} = \left[\frac{k+1}{k}\right]\frac{1}{2(1-\alpha)}. \tag{34}$$

Substituting $(k/k + \alpha)^k$ for $\sigma_k(\alpha)$ in (33) we obtain the following upper bound for the conditional average delay of the observed delayed calls (this bound, as we know, is reached when the observed calls are served last):

$$\frac{[k/(k+\alpha)]^k + \alpha - 1}{\{1 - [k/(k+\alpha)]^k\}\alpha(1-\alpha)}. \tag{35}$$

Subtracting (34) from (35) we find that the difference is of the same sign as

$$\alpha - \{1 - [k/(k+\alpha)]^k\}[1 + (k+1)\alpha/2]. \tag{36}$$

The two factors in the second term of (36) are both increasing functions of k and since (36) vanishes for $k = 1$ [a fact that we already know from (32)] we may conclude that (35) is smaller than (34) for $0 < k < 1$, and greater than (34) for $k > 1$. This proves that, for $k < 1$, the conditional average delay for all delayed calls is still larger than the conditional average delay for all observed delayed calls even if these calls are served last. For $k > 1$, the conditional average delay of the observed delayed calls for the observed-served-last procedure is larger than the conditional average delay of all the delayed calls.

Expressions for the higher moments of the observed calls delay-distribution are also readily obtained. Let F be the equilibrium cumulative distribution of the virtual waiting time, V, at the conclusion of the measurement of a delay (this delay, of course, may be equal to zero). The delay distribution, K, of the calls whose delays are observed, can be readily expressed in terms of F. Indeed we have:

$$K(w) = \Pr[W_* \leqq w] = \alpha \int_0^w \left\{ \int_t^\infty \exp - \alpha(y - t) \cdot dF(y) \right\} dt + \int_0^\infty e^{-\alpha y} dF(y). \tag{37}$$

TABLE I—MEANS AND STANDARD DEVIATIONS OF THE DELAY DISTRIBUTIONS FOR ALL CALLS AND FOR ALL OBSERVED CALLS (1 CLOCK) IN M/M/1—FIRST-COME FIRST-SERVED

α	EW_1	SW_1	EW_{*1}	SW_{*1}
0.1	0.11111	0.48432	0.09259	0.42264
0.2	0.25000	0.75000	0.17840	0.58463
0.3	0.42857	1.0202	0.26684	0.72360
0.4	0.66667	1.3333	0.36669	0.87308
0.5	1.0000	1.7321	0.48958	1.0590
0.6	1.5000	2.2913	0.65554	1.3188
0.7	2.3333	3.1798	0.90754	1.7310
0.8	4.0000	4.8990	1.3659	2.5191
0.9	9.0000	9.9499	2.5808	4.7519
0.92	11.500	12.460	3.1398	5.8266
0.94	15.667	16.637	4.0284	7.5793
0.96	24.000	24.980	5.6974	10.987
0.98	49.000	49.990	10.243	20.776
0.99	99.000	99.995	18.393	39.434

The problem of finding the distribution K of the observed delays is thus reduced to the problem of finding F. The distribution F satisfies the following integral equation:

$$F(t) = \sum_{n=1}^{\infty} \int_0^t dS^{(n)}(u) \int_0^{\infty} e^{-\alpha y} \frac{(\alpha y)^n}{n!} dF(y) + \int_0^t dS(u) \int_0^{\infty} e^{-\alpha v} dF(y), \quad (38)$$

where $S^{(n)}$ stands for the nth convolution of S with itself. Equation (38) follows immediately upon noticing that, at the conclusion of a measurement, the only calls in the system are (*i*) the call whose delay has just been measured and (*ii*) all the calls which arrived during the measurement interval (we note that the preceding argument makes essential use of the assumption that calls are served in order of arrival).

Let φ and σ be respectively the Laplace-Stieltjes transform of F and S. Then transforming (38) we obtain

$$\varphi(s) = \sum_{n=1}^{\infty} \sigma^n(s) \int_0^{\infty} e^{-\alpha y} \frac{(\alpha y)^n}{n!} dF(y) + \sigma(s) \int_0^{\infty} e^{-\alpha v} dF(y) = \varphi\{\alpha[1 - \sigma(s)]\} + \Phi(\alpha)[\sigma(s) - 1]. \quad (39)$$

This formula may be used to derive recurrence relations for the moments of V. Let $n!\mu_n = EV^n$ and

$$n!\nu_n = (-1)^n \frac{d}{ds^n} \sigma(s)\Big|_{s=0} = \int_0^{\infty} t^n dS(t),$$

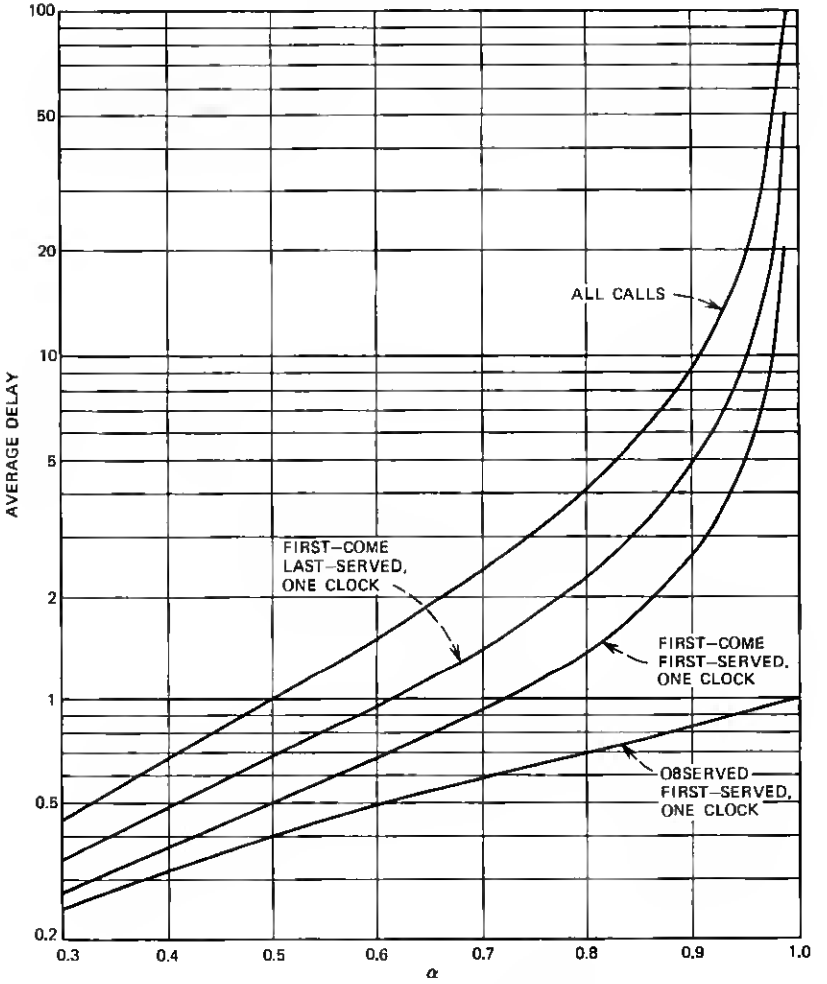

Fig. 1—Average delay for M/M/1 vs occupancy.

so that $n!\nu_n$ is the nth moment of the service-time distribution. Using Faa di Bruno's formula for the derivative of a composite function (Ref. 9, p. 36) we find that:

$$\mu_n = \sum \frac{k!}{k_1! \cdots k_n!} \mu_k \alpha^k \nu_1^{k_1} \cdots \nu_n^{k_n} + \Phi(\alpha)\nu_n, \tag{40}$$

with $k = k_1 + \cdots + k_n$ and the sum over all solutions in non-negative

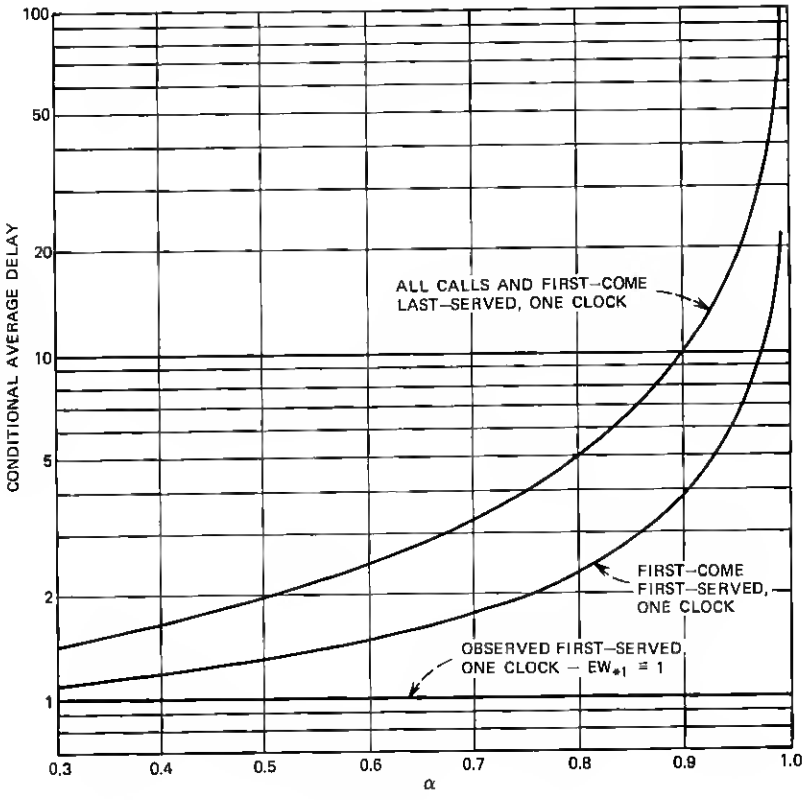

Fig. 2—Conditional average delay for M/M/1 vs occupancy.

integers of $k_1 + 2k_2 + \cdots + nk_n = n$ (note that $\nu_1 = 1$ since the average service-time is assumed here to be equal to 1).

In particular, for $n = 1, 2$, and 3 we have:

$$\mu_1 = EV = \Phi(\alpha)/(1 - \alpha),$$

$$\mu_2 = \frac{EV^2}{2!} = \Phi(\alpha)\nu_2/(1 - \alpha)(1 - \alpha^2),$$

$$\mu_3 = \frac{EV^3}{3!} = \Phi(\alpha)[2\alpha^2\nu_2^2 + \nu_3(1 - \alpha^2)]/(1 - \alpha)(1 - \alpha^2)(1 - \alpha^3),$$

where $\Phi(\alpha)$ is the probability that an observed call is not delayed.

When the service-time distribution is exponential (with mean 1) we have $\nu_i = 1, i = 0, 1, \cdots$, and (40) becomes:

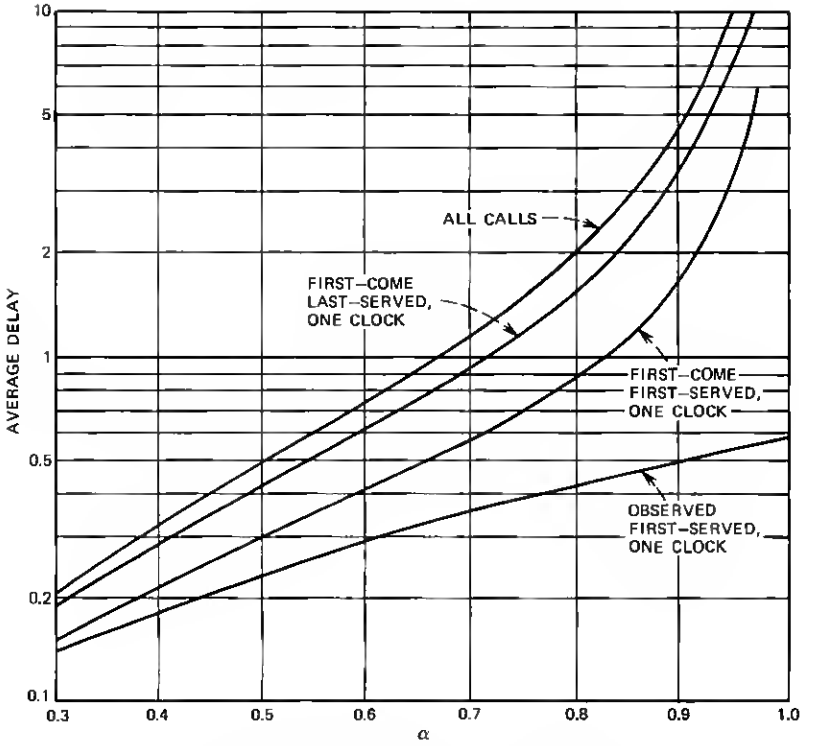

Fig. 3—Average delay for M/D/1 vs occupancy.

$$\mu_n = \sum \frac{k!}{k_1! \cdots k_n!} \mu_k \alpha^k + \Phi_1(\alpha)$$

$$= \sum_{k=1}^{n} \binom{n-1}{k-1} \mu_k \alpha^k + \Phi_1(\alpha), \qquad n \geqq 1.$$

Equation (37) may be used to express the moments of K in terms of the moments of V. We have:

$$EW_*^n = \int_0^\infty w^n dK(w) = \alpha \int_0^\infty w^n \int_w^\infty \exp - \alpha(y - w) dF(y) dw,$$

and, upon integrating by parts, we obtain

$$EW_* = EV - \frac{1}{\alpha}[1 - \Phi(\alpha)],$$

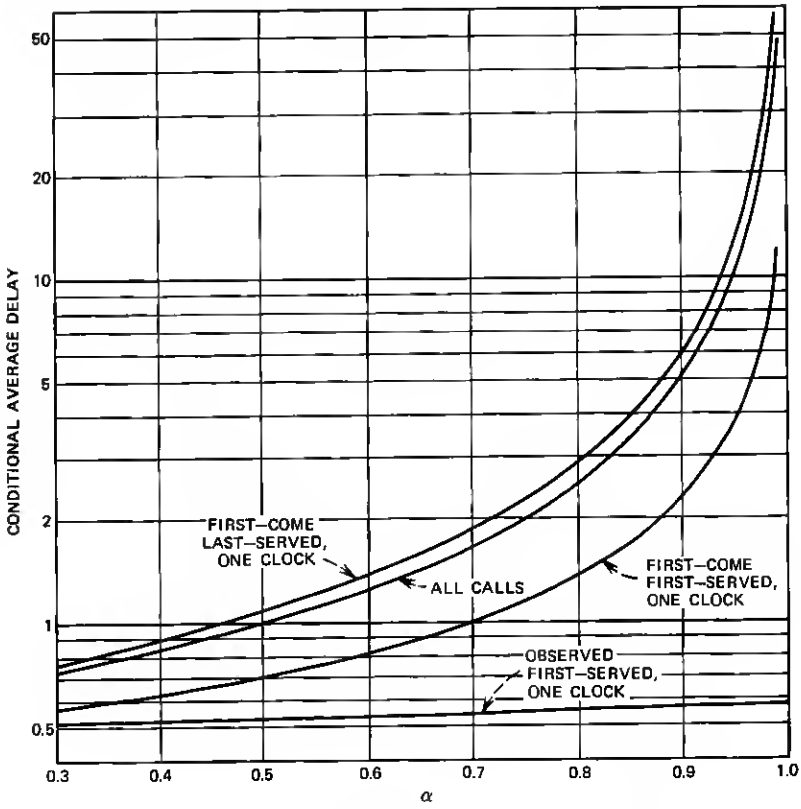

Fig. 4—Conditional average delay for M/D/1 vs occupancy.

and

$$EW_*^{n+1} = EV^{n+1} - \frac{n+1}{\alpha} EW_*^n, \qquad n > 0.$$

Thus, in particular, we have

$$EW_* = \frac{\Phi(\alpha) + \alpha - 1}{\alpha(1-\alpha)},$$

$$EW_*^2 = \frac{2\alpha^2\Phi(\alpha)\nu_2 - 2[\Phi(\alpha) + \alpha - 1](1-\alpha^2)}{\alpha^2(1-\alpha)(1-\alpha^2)}.$$

The moments of W_* depend only on the moments of the service-time distribution and on $\Phi(\alpha)$.

As a numerical illustration of the biases induced when only one clock is available, the means and the standard deviations of W_1 and

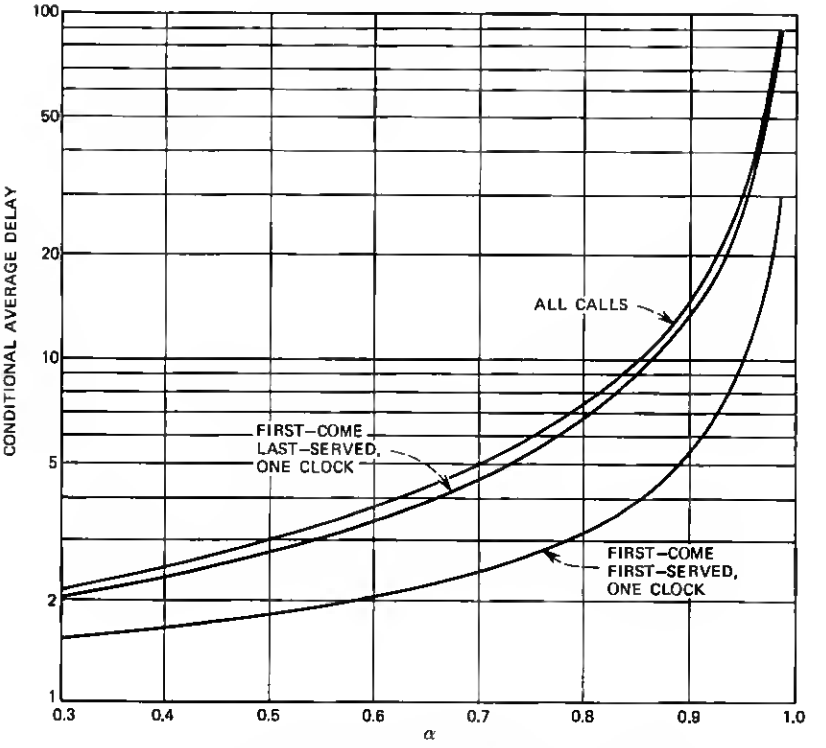

Fig. 5—Conditional average delay for $M/\Gamma_{\frac{1}{2}}/1$ vs occupancy.

W_{*1} are given in Table I. (The standard deviations of W_1 and W_{*1} are designated by SW_1 and SW_{*1} respectively.) For further quantitative results, see Figs. 1–6.

VI. THE SINGLE-SERVER QUEUE M/M/1

In this section we consider a single-server delay system and assume that: (*i*) calls arrive in a Poisson process of intensity α; (*ii*) the service-times are independent random variables with the same negative exponential distribution; and (*iii*) calls are served in order of arrival. We again suppose that only one delay can be measured at a time. Our purpose here is to derive the delay distribution of the observed calls.

Let $F(\cdot)$ be the equilibrium cumulative distribution of the virtual waiting time at the conclusion of the measurement of a delay; the measured delay may of course be equal to zero. From (38), the distri-

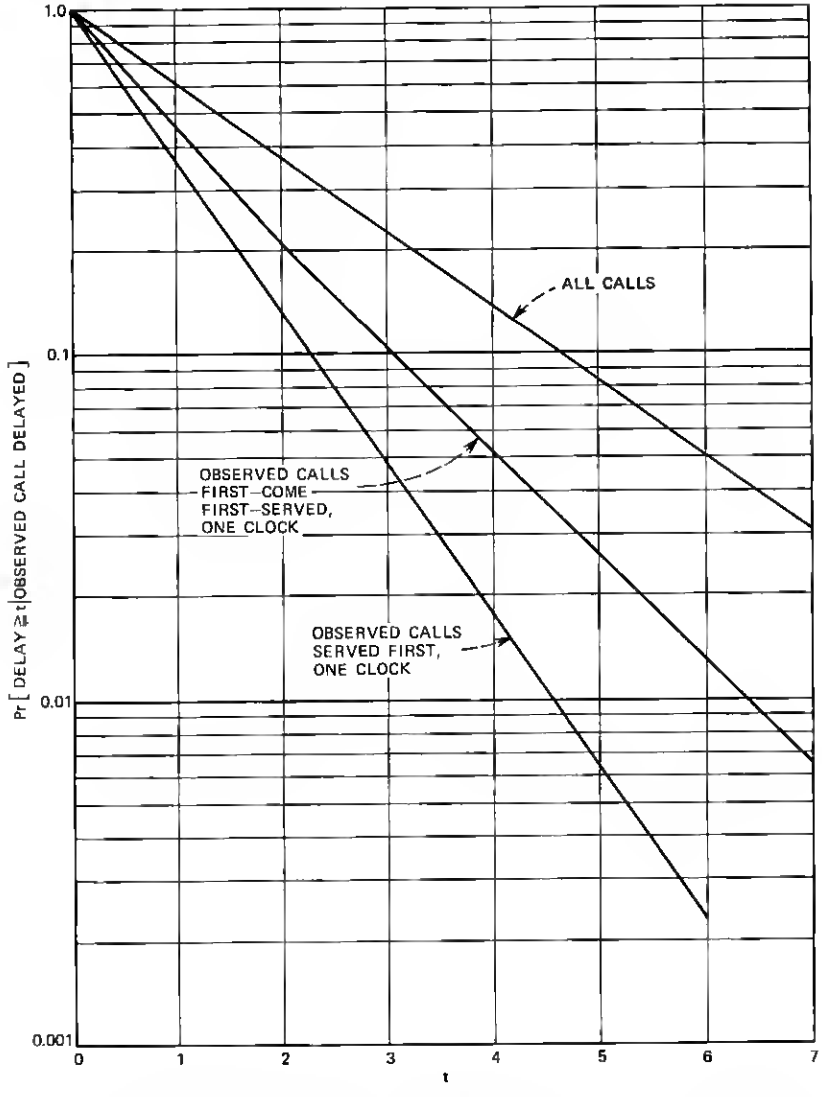

Fig. 6—Conditional delay distributions for M/M/1 — $\alpha = 0.5$ first-come first-served.

bution $F(\cdot)$ satisfies tbe following integral equation:

$$F(t) = \int_0^t \left\{ \int_0^\infty \sum_{n=1}^\infty e^{-\alpha y} \frac{(\alpha y)^n}{n!} \frac{u^{n-1}}{(n-1)!} e^{-u} dF(y) \right\} du + \int_0^t e^{-u} du \cdot \int_0^\infty e^{-\alpha y} dF(y).$$

This relation implies that $F(\cdot)$ is continuous and that the virtual waiting time, at the conclusion of a measurement, has a density function $f(\cdot)$ [at $t = 0$, the latter is defined as the right-hand derivative of $F(\cdot)$]. We have therefore

$$f(t) = e^{-t}\int_0^\infty f(y)e^{-\alpha y}\sum_{n=1}^{\infty}\frac{(\alpha y)^n}{n!}\frac{t^{n-1}}{(n-1)!}\cdot dy + e^{-t}\int_0^\infty f(y)e^{-\alpha y}dy$$

$$= \alpha^{\frac{1}{2}}e^{-t}t^{-\frac{1}{2}}\int_0^\infty f(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy + e^{-t}\int_0^\infty f(y)e^{-\alpha y}dy, \qquad (41)$$

where $I_1(\cdot)$ is the modified Bessel function of order 1 (Ref. 7, p. 374).

The preceding relation implies that $f(\cdot)$ is of the form

$$f(t) = f_1(t) + ce^{-t},$$

where c is a constant. Substitution of this expression in (41) yields, on taking relation 29.3.81, p. 1026, of Ref. 7 into account:

$$f_1(t) = \alpha^{\frac{1}{2}}e^{-t}t^{\frac{1}{2}}\int_0^\infty f_1(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy + \frac{c\alpha e^{-t}}{(1+\alpha)^2}e^{\alpha t/(1+\alpha)}. \qquad (42)$$

Thus $f_1(\cdot)$ is of the form

$$f_1(t) = f_2(t) + \frac{c\alpha}{(1+\alpha)^2}\exp - t/(1+\alpha),$$

and substituting this expression in (42) we find that $f_2(\cdot)$ is of the form

$$f_2(t) = f_3(t) + \frac{c\alpha^2}{(1+\alpha+\alpha^2)^2}\exp - t/(1+\alpha+\alpha^2).$$

Proceeding in this manner, we define successively $f_4(\cdot)$, $f_5(\cdot)$, $\cdots$, and it is readily shown, by induction, that:

$$f_m(t) = f_{m+1}(t) + \frac{c\alpha^m}{(1+\alpha+\alpha^2+\cdots+\alpha^m)^2}$$
$$\cdot\exp - t/(1+\alpha+\cdots+\alpha^m), \qquad m = 0, 1, 2, \cdots;$$
$$f_0(\cdot) = f(\cdot). \qquad (43)$$

Passing to the limit, we obtain, in this manner:

$$f(t) = f_\infty(t) + c\sum_{m=0}^{\infty}\frac{\alpha^m}{(1+\alpha+\cdots+\alpha^m)^2}$$
$$\cdot\exp - t/(1+\alpha+\cdots+\alpha^m),$$

where $f_\infty(\cdot) = \lim_{m\to\infty} f_m(\cdot)$ satisfies the integral equation:

$$f_\infty(t) = \alpha^{\frac{1}{2}}e^{-t}t^{-\frac{1}{2}}\int_0^\infty f_\infty(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy. \qquad (44)$$

Note that, by virtue of (43), $f_m(t) > f_{m+1}(t)$ for all t and that $\lim_{m\to\infty} f_m(t)$ does therefore exist and is non-negative since $f_m > 0$ for all m.

We shall prove now that the only non-negative solution of (44) is $f_\infty(t) \equiv 0$.

Let

$$\theta(s) \equiv \int_0^\infty f_\infty(t)e^{-st}dt.$$

Then, transforming the previous relation, we have by Ref. 7, p. 1026, equation 29.3.81:

$$\begin{aligned}\theta(s) &= \alpha^{\frac{1}{2}} \int_0^\infty e^{-st}e^{-t}t^{-\frac{1}{2}} \int_0^\infty f_\infty(y)e^{-\alpha y}y^{\frac{1}{2}}I_1[2(\alpha yt)^{\frac{1}{2}}]dy \cdot dt \\ &= \alpha^{\frac{1}{2}} \int_0^\infty f_\infty(y)e^{-\alpha y}y^{\frac{1}{2}}(\alpha y)^{-\frac{1}{2}}(e^{\alpha y/1+s} - 1)dy \\ &= \int_0^\infty f_\infty(y) \exp - \alpha y\left(1 - \frac{1}{1+s}\right)dy \\ &\qquad - \int_0^\infty f_\infty(y) \exp(-\alpha y)dy \\ &= \theta\left(\frac{\alpha s}{1+s}\right) - \theta(\alpha).\end{aligned}$$

Setting s equal to zero in the preceding relation, we find that $\theta(\alpha) = 0$ which implies that $f_\infty(t) \equiv 0$ and we have, therefore, for exponential service-times:

$$\begin{aligned} f(t) &= c \sum_{m=0}^{\infty} \frac{\alpha^m}{(1+\alpha+\cdots+\alpha^m)^2} \\ &\qquad \cdot \exp - t/(1+\alpha+\cdots+\alpha^m), \qquad t \geqq 0, \\ 1 - F(t) &= c \sum_{m=0}^{\infty} \frac{\alpha^m}{1+\alpha+\cdots+\alpha^m} \\ &\qquad \cdot \exp - t/(1+\alpha+\cdots+\alpha^m), \qquad t \geqq 0, \\ f(t) &= F(t) = 0, \qquad t < 0, \end{aligned} \tag{45}$$

where the constant c is determined by the condition $F(0) = 0$.

By means of (37) and (45), it is readily seen that the conditional delay distribution is given by the following formula:

$$\begin{aligned}\Pr[W_* \geqq t | \text{observed call delayed}] &= c' \sum_{m=0}^{\infty} \frac{\alpha^{m+1}}{1+\alpha+\cdots+\alpha^{m+1}} \\ &\qquad \cdot \exp - t/(1+\alpha+\cdots+\alpha^m), \qquad t > 0,\end{aligned}$$

where c' is determined hy the requirement that the preceding expression be equal to 1 for $t = 0$.

The effect of partial sampling on the delay distribution is illustrated in Fig. 6.

VII. THE MULTISERVER QUEUE M/M/s

In this section we consider a full-access multiserver delay system with Poisson arrivals and exponential service-times. Our purpose here is to determine the probability $\Phi_1^{(s)}$ that an observed call is served without delay and the expected delay $EW_{*1}^{(s)}$ of the observed calls ($\Phi_1^{(1)} = \Phi_1$, $W_{*1}^{(1)} = W_{*1}$). This is easily done. Indeed, under the present assumptions, $A_1^{(s)}$ the expected number of nonobserved calls during the measurement of a positive delay is:

$$A_1^{(s)} = \frac{\alpha EW_{*1}}{s(1 - \Phi_1)}, \tag{46}$$

where EW_{*1} and Φ_1 pertain to the single-server queue with load α/s. Equation (46) is an immediate consequence of the fact that in an s-server system with demand rate α and service rate 1, the conditional average delay of an observed call, $EW_{*1}^{(s)}/(1 - \Phi_1^{(s)})$, is equal to the conditional average delay of an observed call in a single-server queue with offered load α/s and service rate s, i.e., $EW_{*1}/s(1 - \Phi_1)$.

Hence, by (46) and (1) we have

$$\Phi_1^{(s)} = \frac{(1 - B)[(\alpha/s)EW_{*1} + 1 - \Phi_1]}{(1 - B)[(\alpha/s)EW_{*1} + 1 - \Phi_1] + B(1 - \Phi_1)}$$

so that, by (17),

$$EW_{*1}^{(s)} = \frac{B \cdot EW_{*1}}{s(1 - \Phi_1) + \alpha(1 - B)EW_{*1}}.$$

We note that

$$\frac{EW_{*1}^{(s)}}{EW_1^{(s)}} = \frac{(1 - \alpha/s)EW_{*1}}{1 - \Phi_1 + (1 - B)(\alpha/s)EW_{*1}}.$$

For α/s fixed, the blocking prohahility B is strictly decreasing and tends to 0 as s increases. Hence

$$\frac{EW_{*1}^{(s)}}{EW_1^{(s)}} > \frac{EW_{*1}^{(s+m)}}{EW_1^{(s+m)}}, \qquad m > 0,$$

and

$$\left[\frac{EW_*}{EW}\right]_\infty \equiv \lim_{s \to \infty} \frac{EW_{*1}^{(s)}}{EW_1^{(s)}} = \frac{(1 - \alpha/s)EW_{*1}}{1 - \Phi_1 + (\alpha/s)EW_{*1}}.$$

We stress that the preceding relations are valid regardless of the order of service.

Numerical values are given in Table II. They show, in particular, that, for a given server occupancy, the magnitudes of the relative biases become larger as the number of servers, s, increases but remain bounded.

VIII. AN INEQUALITY FOR GI/M/s

For the M/G/1 queue we have seen that the average delay on all calls, EW, is always larger than the expected delay EW_* even if the observed calls are served last. It will be shown here that the same relation also holds for the multiserver queue GI/M/s.

When the observed calls are served last, the waiting times of the observed delayed calls have the same distribution as the busy period whenever the service-times are exponential. Writing $\overline{EW}_{*1}^{(s)}$ for the unconditional average delay for the observed-served-last procedure we have therefore:

$$\overline{EW}_{*1}^{(s)} = (1 - \Phi)/s(1 - b), \tag{47}$$

where b is the root of smallest absolute value of the equation (Ref. 10, p. 225 ff.)

$$z = \mathbf{A}^*(1 - z)$$

and $\mathbf{A}^*$ is the Laplace-Stieltjes transform of the interarrival distribution $\mathbf{A}$. We note here that b is also the blocking probability in the associated single-server queue GI/M/1 with $\mathbf{A}$ as interarrival distribution and exponential service-time distribution with mean $1/s$.

By means of (1) we can rewrite (47) as follows:

$$\overline{EW}_{*1}^{(s)} = B/s(1 - b)[(1 - B)(1 + A) + B], \tag{48}$$

where B is the probability of delay in the GI/M/s queue.

When the observed calls are served last, $1 + A$ is equal to the expected number of calls served during a busy period of GI/M/s which, in turn, is equal to the expected number of calls served during a busy period of the associated single-server queue GI/M/1. Hence, we have (Ref. 10, p. 286):

$$1 + A = (1 - b)^{-1},$$

and on taking this relation into account, (48) reduces to

$$\overline{EW}_{*1}^{(s)} = B/s[1 - B + B(1 - b)].$$

TABLE II—AVERAGE DELAYS IN M/M/s—FIRST-COME FIRST-SERVED

α/s	EW_{*1}	EW_{*1}/EW_1	$EW_{*1}^{(2)}$	$EW_{*1}^{(2)}/EW_1^{(2)}$	$EW_{*1}^{(4)}$	$EW_{*1}^{(4)}/EW_1^{(4)}$	$EW_{*1}^{(8)}$	$EW_{*1}^{(8)}/EW_1^{(8)}$	$[EW_{*1}/EW_1]_\infty$
0.1	0.093	0.83	0.008	0.82					0.82
0.2	0.178	0.71	0.029	0.70	0.002	0.69			0.69
0.3	0.227	0.62	0.059	0.60	0.008	0.58	0.0004	0.57	0.57
0.4	0.337	0.55	0.099	0.52	0.018	0.49	0.0019	0.48	0.48
0.5	0.490	0.49	0.151	0.45	0.037	0.42	0.0059	0.40	0.39
0.6	0.656	0.44	0.224	0.40	0.065	0.36	0.0146	0.34	0.31
0.7	0.908	0.39	0.336	0.35	0.111	0.31	0.0316	0.28	0.24
0.8	1.37	0.34	0.541	0.30	0.199	0.27	0.0665	0.23	0.16
0.9	2.58	0.29	1.09	0.26	0.438	0.22	0.166	0.19	0.086
0.95	4.71	0.25	2.06	0.22	0.866	0.19	0.349	0.17	0.045

But the average delay for all calls is given by the formula (Ref. 11, p. 383):

$$EW_1^{(s)} = B/s(1 - b)$$

so that

$$\overline{EW}_{\neq 1}^{(s)} < EW_1^{(s)}.$$

REFERENCES

1. *Proceedings of the Symposium on Congestion Theory*, W. L. Smith and W. E. Wilkinson, eds., Chapel Hill, N. C.: The University of North Carolina Press, 1965.
2. Oberer, E., and Riesz, G. W., "Test Calling: Experience, Theory, Prospect," Proc. Seventh Int. Teletraffic Congress, Stockholm, June 13–20, 1973.
3. Riordan, J., *Stochastic Service Systems*, New York: Wiley, 1962.
4. Loève, M., *Probability Theory*, third edition, Princeton, N. J.: D. Van Nostrand, 1963.
5. Doob, J. L., *Stochastic Processes*, New York: Wiley, 1953.
6. Smith, W. L., "Renewal Theory and Its Ramifications," J. Roy. Stat. Soc., Ser. B, *20*, No. 2, 1958, pp. 243–302.
7. *Handbook of Mathematical Functions*, M. Abramowitz and I. A. Stegun, eds., National Bureau of Standards, Applied Mathematics Series, 55, 1965.
8. Takács, L., *Introduction to the Theory of Queues*, New York: Oxford University Press, 1962.
9. Riordan, J., *An Introduction to Combinatorial Analysis*, New York: Wiley, 1958.
10. Cohen, J. W., *The Single-Server Queue*, New York: American Elsevier Publishing Company, Inc., 1969.
11. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Edinburgh and London: Oliver and Boyd, 1960.